Nucleic Acids Research, 2003, Vol. 31, No. 1 **187–189**
DOI: 10.1093/nar/gkg004

HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes

S. Garcia-Vallve*, E. Guzman, M. A. Montero and A. Romeu

Evolutionary Genomics Group, Biochemistry and Biotechnology Department, 'Rovira i Virgili' University,
Pl Imperial Tàrraco 1, E-43005 Tarragona, Spain

Received July 17, 2002; Revised and Accepted September 12, 2002

ABSTRACT

The Horizontal Gene Transfer DataBase (HGT-DB) is a genomic database that includes statistical parameters such as G+C content, codon and amino-acid usage, as well as information about which genes deviate in these parameters for prokaryotic complete genomes. Under the hypothesis that genes from distantly related species have different nucleotide compositions, these deviated genes may have been acquired by horizontal gene transfer. The current version of the database contains 88 bacterial and archaeal complete genomes, including multiple chromosomes and strains. For each genome, the database provides statistical parameters for all the genes, as well as averages and standard deviations of G+C content, codon usage, relative synonymous codon usage and amino-acid content. It also provides information about correspondence analyses of the codon usage, plus lists of extraneous group of genes in terms of G+C content and lists of putatively acquired genes. With this information, researchers can explore the G+C content and codon usage of a gene when they find incongruities in sequence-based phylogenetic trees. A search engine that allows searches for gene names or keywords for a specific organism is also available. HGT-DB is freely accessible at http://www.fut.es/~debb/HGT.

INTRODUCTION

Horizontal Gene Transfer (HGT), the transfer of genes between different species, is recognized as one of the major forces in prokaryotic genome evolution (1). Acquired genes may provide novel metabolic capabilities and catalyze the diversification of microbial lineages. HGT events can be detected from patterns of best matches to different species and the distribution of genes, or by identifying regions of the genome with unusual compositions or incongruities between phylogenetic trees (2,3). Each of these methods has its advantages and disadvantages (2). The prediction of horizontally transferred genes using atypical nucleotide composition is based on the genome hypothesis (4) that assumes that codon usage and G+C content are distinct global features of each prokaryotic genome. With this method, a significant number of prokaryotic genes have been proposed as having been acquired by HGT (5,6). However, it cannot predict all acquired genes unambiguously (7) because genes may have adjusted to the base composition and codon usage of the host genome (this is called the amelioration process) or because an unusual composition may be due to factors other than HGT (6). Despite these limitations, atypical G+C content and patterns of codon usage are especially useful for detecting the putative origin of the transferred genes (8–10).

To confirm whether a gene or group of genes has been acquired by HGT, it can be useful to combine multiple lines of evidence (2). If researchers have access to the compositional parameters for each gene from complete genomes, they will be able to explore for themselves the G+C content and codon usage of genes when they find incongruences among sequence-based phylogenetic trees or when they detect putatively transferred genes with other methods. We have, therefore, created the Horizontal Gene Transfer DataBase (HGT-DB) to facilitate compositional analyses and provide additional evidence for discussing the possible foreign origin of the genes of a genome and detecting whether acquired genes have been ameliorated. For each prokaryotic complete genome, the HGT-DB provides averages and standard deviations of G+C content, codon usage, relative synonymous codon usage and amino-acid content, as well as lists of putative horizontally transferred genes, correspondence analyses of the codon usage and lists of extraneous groups of genes in terms of G+C content. For each gene, the database lists several statistical parameters, including total and positional G+C content, and determines whether the gene deviates from the mean values of its own genome. The HGT-DB has so far been used to study strain-specific genes of *Helicobacter pylori* (11,12) and to exclude putative horizontally transferred genes in genomic or proteomic analyses (13).

SOURCES OF GENOMIC DATA AND METHODS

Sequence files of prokaryotic complete genomes are retrieved from the NCBI ftp server. Total and positional G+C content, codon usage, relative synonymous codon usage and amino-acid content are calculated for each gene. For each genome,

*To whom correspondence should be addressed. Tel: +34 977559565; Fax: +34 977558232; Email: vallve@quimica.urv.es

Table 1. Species, total number of Open Reading Frames (ORF) and number (*N*) and percentage (%) of extraneous genes in terms of G+C content and codon usage from archaeal and bacterial complete genomes included in the database

Genome	ORF	*N*	%
Archaea			
Aeropyrum pernix K1	1840	270	15.7
Archaeoglobus fulgidus	2420	160	7.7
Halobacterium sp. NRC-1	2075	149	8.4
Methanobacterium thermoautotrophicum deltaH	1873	178	10.9
Methanococcus jannaschii	1729	72	4.8
Methanopyrus kandleri AV19	1687	179	11.5
Methanosarcina acetivorans	4540	602	15.1
Methanosarcina mazei	3371	378	12.6
Pyrobaculum aerophilum	2605	308	14.5
Pyrococcus abyssi	1769	121	7.3
Pyrococcus furiosis	2065	134	7.4
Pyrococcus horikoshii	1801	123	7.3
Sulfolobus solfataricus	2977	147	5.4
Sulfolobus tokodaii	2826	132	5.2
Thermoplasma acidophilum	1482	145	10.8
Thermoplasma volcanium	1499	104	7.8
Bacteria			
Agrobacterium tumefaciens str. C58 (Cereon) circular chromosome	2721	194	7.6
Agrobacterium tumefaciens str. C58 (Cereon) linear chr.	1833	114	6.5
Agrobacterium tumefaciens str. C58 (U. Wash.) circular chr.	2785	142	5.7
Agrobacterium tumefaciens str. C58 (U. Wash.) linear chr.	1876	114	6.5
Aquifex aeolicus	1529	70	4.8
Bacillus halodurans C-125	4066	304	8.6
Bacillus subtilis	4112	552	15.0
Borrelia burgdorferi	851	10	1.4
Brucella melitensis chr. I	2059	118	6.5
Brucella melitensis chr. II	1139	59	5.7
Buchnera aphidicola Sg	544	6	1.3
Buchnera sp. APS	564	0	0.0
Campylobacter jejuni	1634	78	5.4
Caulobacter crescentus	3737	135	3.9
Chlorobium tepidum TLS	2252	267	14.5
Chlamydophila pneumoniae J138	1069	49	5.2
Chlamydophila pneumoniae CWL029	1054	58	6.0
Chlamydophila pneumoniae AR39	1112	55	5.9
Chlamydia trachomatis	895	36	4.3
Chlamydia muridarum	909	12	1.5
Clostridium acetobutylicum ATCC824	3672	146	4.4
Clostridium perfringens	2660	75	3.2
Corynebacterium glutamicum	3040	207	7.5
Deinococcus radiodurans chr. 1	2629	86	3.5
Deinococcus radiodurans chr. 2	368	23	6.4
Escherichia coli K12	4279	359	9.2
Escherichia coli O157	5361	625	13.3
Escherichia coli O157 : H7 : EDL933	5324	593	12.6
Fusobacterium nucleatum ATCC25586	2067	40	2.2
Haemophilus influenzae Rd	1714	87	5.7
Helicobacter pylori 26695	1576	87	6.3
Helicobacter pylori J99	1491	68	4.9
Lactococcus lactis	2267	90	4.5
Listeria innocua	2968	164	6.2
Listeria monocytogenes EGD-e	2846	184	7.1
Mesorhizobium loti	6746	604	9.9
Mycobacterium leprae TN	1605	73	5.1
Mycobacterium tuberculosis H37Rv	3927	176	4.8
Mycobacterium tuberculosis CDC1551	4187	197	5.4
Mycoplasma genitalium G37	484	51	11.9
Mycoplasma pneumoniae M129	689	39	6.2
Mycoplasma pulmonis UAB CTIP	782	28	4.0
Neisseria meningitidis MC58	2079	221	12.5
Neisseria meningitidis Z2491	2065	206	11.7
Nostoc sp. PCC 7120	5366	203	4.4
Pasteurella multocida PM70	2015	117	6.1
Pseudomonas aeruginosa PA01	5567	307	5.9
Ralstonia solanacearum	3440	356	11.2
Rickettsia conorii Malish 7	1374	54	5.6
Rickettsia prowazekii MadridE	835	28	3.6
Salmonella entereica serovar typhi	4395	551	13.9
Salmonella enterica serovar typhimurium LT2	4451	446	11.0
Sinorhizobium meliloti 1021	3341	179	5.8
Staphylococcus aureus Mu50	2714	119	5.1
Staphylococcus aureus MW2	2632	131	5.8
Staphylococcus aureus N315	2594	105	4.6
Streptococcus pneumonia R6	2043	249	14.1
Streptococcus pneumonia TIGR4	2094	258	15.1
Streptococcus pyogenes SF320	1697	136	9.1
Streptococcus pyogenes MGAS8232	1845	157	10.0
Streptomyces coelicolor A3 (2)	7512	541	7.8
Synechocystis PCC6803	3167	211	7.3
Thermoanaerobacter tengcongensis	2588	343	14.9
Thermotoga maritima	1858	194	11.6
Treponema pallidum subsp. pallidum	1036	78	8.7
Ureaplasma urealyticum	614	12	2.3
Vibrio cholerae chr. 1	2742	234	10.0
Vibrio cholerae chr. 2	1093	204	22.2
Xanthomonas campestris	4181	285	7.4
Xanthomonas citri	4312	284	7.1
Xylella fastidiosa	2766	458	21.4
Yersinia pestis CO92	3885	316	9.0

The percentages are referred to the genes analyzed, that exclude genes smaller than 300 bp and genes for ribosomal proteins.

except for genes under 300 bp, which can have extraneous compositional values, the averages and standard deviations of the above parameters are calculated. The methods we used to consider whether a gene is extraneous in terms of G+C content or codon usage and a candidate to be acquired by HGT are described in Garcia-Vallve *et al.* (6). Briefly, genes are considered as extraneous in terms of G+C content or codon usage if they deviate by more than 1.5 standard deviations from the mean values. Genes are considered to be putative horizontally transferred genes if they have extraneous G+C content and codon usage, they are over 300 bp and they do not deviate from the average amino-acid composition. Clusters of genes with a high or low G+C content are also considered to be acquired genes, regardless of their length or codon usage (6). It is important to distinguish highly expressed genes from horizontally transferred genes (6). Highly expressed genes may deviate from the mean values of codon usage because they adapt their codon usage to the more abundant tRNAs. For this reason, ribosomal proteins, a group of highly expressed genes, are filtered and not included in the database predictions. Other groups of highly expressed genes will be included in future versions of the database, but individual analyses to define the group of highly expressed genes for each genome, if there are any, will probably be needed.

Genes proposed as being acquired horizontally are represented in a correspondence analysis in which protein-coding sequences are considered as points in a 59-dimensional space (the stop codons and codons for methionine and tryptophan are not included), and each dimension corresponds to the relative frequency of use of each codon measured with the relative synonymous codon usage (RSCU) values. Correspondence analysis reduces this multidimensional space to a two- or three-dimensional space that can be represented graphically. In these graphs, vertically descended genes are expected to cluster together around the origin, whereas genes predicted as acquisitions are expected to be on the periphery.

ORGANIZATION OF THE DATABASE

The HGT-DB is organized by genome i.e. every prokaryotic genome that has been completely sequenced forms a new entry. Different chromosomes from the same organism, or genomes from the same species but different strains, are found in different entries. The current version of the database contains 88 genomes that are sorted alphabetically and classified taxonomically. Table 1 shows the archaeal and bacterial genomes included in the current version of the database, as well as the number of extraneous genes in terms of G+C content and codon usage. The main page for each genome contains links to additional sections and the mean values and standard deviations of total and positional G+C content, codon usage, relative synonymous codon usage and amino-acid content. The other sections available for each genome are: a correspondence analysis of the codon usage, a list of extraneous regions in terms of G+C content and a list of the proposed horizontally acquired genes. The database also provides access to a tab-delimited file with all the statistical calculations for each gene of a genome. The fields available for each gene in these files are: information about its position (coordinates, strand and length), gene name, function, the Cluster of Orthologous Group, COG, (14) it belongs to, total and positional G+C content, the Mahalanobis distance to the average codon usage (6), amino-acid content deviations, if any, and a prediction of whether the gene belongs to a region with a high or low G+C content or whether it has been acquired by HGT. This information can be also accessed via a search engine that allows searches for gene names or keywords for a specific organism. When searching for a gene name, one can also view the upstream and downstream genes.

Forces other than HGT are also responsible for the heterogeneity in the codon usage of all the genes of a genome. The HGT-DB, therefore, has a section containing the correspondence analysis of the relative synonymous codon usage for each genome. This section contains a table with the percentage variability of the six axes that account for the greatest variation in codon usage, a graphical representation of the coordinates of each gene in the first and second axes (the genes proposed as being acquired by HGT and putative highly expressed genes are shown in different colors) and a table with the correlation values between the position of genes in the first or second axis, and the G+C content and several indices of codon bias. These indices are: the effective number of codons (Nc) (15), the intrinsic codon deviation index (ICDI) (16), the translational efficiency index (P2) (17) and the scaled X^2 index (18).

DATABASE ACCESS

HGT-DB is freely accessible at http://www.fut.es/~debb/HGT/. The database will be updated several times each year. Changes and new additions to the database can be viewed in the 'news and previous release' section.

ACKNOWLEDGEMENTS

We thank Kevin Costello of the Language Service of the Rovira i Virgili University for his help with writing the manuscript, and TINET for hosting the database.

REFERENCES

1. Koonin,E.V., Makarova,K.S. and Aravind,L. (2001) Horizontal gene transfer in prokaryotes: Quantification and Classification. *Annu. Rev. Microbiol.*, **55**, 709–742.
2. Eisen,J.A. (2000) Horizontal gene transfer among microbial genomes: new insights from complete genome analysis. *Curr. Opin. Genet. Dev.*, **10**, 606–611.
3. Ragan,M.A. (2001) Detection of lateral gene transfer among microbial genomes. *Curr. Opin. Genet. Dev.*, **11**, 620–626.
4. Grantham,R., Gautier,C., Gouy,M., Mercier,R. and Pave,A. (1980) Codon catalog usage and the genome hypothesis. *Nucleic Acids Res.*, **8**, r49–r62.
5. Ochman,H., Lawrence,J.G. and Groisman,E.A. (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature*, **405**, 299–304.
6. Garcia-Vallve,S., Romeu,A. and Palau,J. (2000) Horizontal gene transfer in bacterial and archaeal complete genomes. *Genome Res.*, **10**, 1719–1725.
7. Lawrence,J.G. and Ochman,H. (2002) Reconciling the many faces of lateral gene transfer. *Trends Microbiol.*, **10**, 1–4.
8. Garcia-Vallve,S., Palau,J. and Romeu,A. (1999) Horizontal gene transfer in glycosyl hydrolases inferred from codon usage in *Escherichia coli* and *Bacillus subtilis. Mol. Biol. Evol.*, **16**, 1125–1134.
9. Garcia-Vallve,S., Romeu,A. and Palau,J. (2000) Horizontal gene transfer of glycosyl hydrolases of the rumen fungi. *Mol. Biol. Evol.*, **17**, 352–361.
10. Garcia-Vallve,S., Simó,F.X., Montero,M.A., Arola,L. and Romeu,A. (2002) Simultaneous horizontal gene transfer of a gene coding for ribosomal protein L27 and operational genes in *Arthrobacter* sp. *J. Mol. Evol.*, in press.
11. Israel,D.A., Salama,N., Krishna,U., Rieger,U.M., Atherton,J.C., Falkow,S. and Peek,R.M.,Jr (2001) *Helicobacter pylori* genetic diversity within the gastric niche of a single human host. *Proc. Natl Acad. Sci. USA*, **98**, 14625–14630.
12. Garcia-Vallve,S., Janssen,P.J. and Ouzounis,C.A. (2002) Genetic variation between *Helicobacter pylori* strains: gene acquisition or loss? *Trends Microbiol.*, **10**, 445–447.
13. Akashi,H. and Gojobori,T. (2002) Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. *Proc. Natl Acad. Sci. USA*, **99**, 3695–3700.
14. Tatusov,R.L., Natale,D.A., Garkavtsev,I.V., Tatusova,T.A., Shankavaram,U.T., Rao,B.S., Kiryutin,B., Galperin,M.Y., Fedorova,N.D. and Koonin,E.V. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.*, **29**, 22–38.
15. Wright,F. (1990) The effective number of codons used in a gene. *Gene*, **87**, 23–29.
16. Freire-Picos,M.A., Gonzalez-Siso,M.I., Rodriguez-Belmonte,E., Rodriguez-Torres,A.M., Ramil,E. and Cerdan,M.E. (1994) Codon usage in *Kluyveromyces lactis* and in yeast cytochrome c-encoding genes. *Gene*, **139**, 43–49.
17. Gouy,M. and Gautier,C. (1982) Codon usage in bacteria-correlation with gene expressivity. *Nucleic Acids Res.*, **10**, 7055–7074.
18. Shields,D.C. and Sharp,P.M. (1987) Synonymous codon usage in *Bacillus subtilis* reflects both translational selection and mutational biases. *Nucleic Acids Res.*, **15**, 8023–8040.